

Performance Metrics

Reading:

Leinwand, A. & Conroy, K. F. (1996) Network Management: A Practical Perspective 2nd ed. Addison-Wesley. Chapters 6.

Stallings, W. (1999) SNMP, SNMPv2, SNMPv3, and RMON 1 and 2. 3rd. ed. Reading Massachusetts, Addison-Wesley. Chapters 2(section 2.2).

rfcs: rfc1857.

Some useful URLs http://www-staff.it.uts.edu.au/~akmzrahm/reports/report4_draft.htm

Performance Indicators

A system or activity cannot be managed or controlled unless its performance is monitored. Measures of performance can be classified as either service-oriented or efficiency-oriented.

Service-oriented Performance Indicators

- Availability: The fraction of the time that a network system, component or an application is available to the user.
- Response time: The time it takes for the system to response to a request to perform a particular task.
- Accuracy : The percentage of time that no errors appear in the transmission and delivery of information.

Availability

This is based on the reliability of the individual components of a network.

Reliability: The probability that a component will perform its specified function for a specified time under specified conditions.

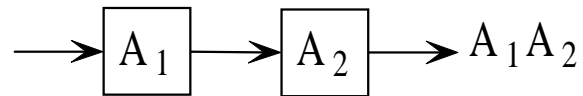
For a component the availability can be expressed as

$$A = \frac{MTBF}{MTBF + MTTR}$$

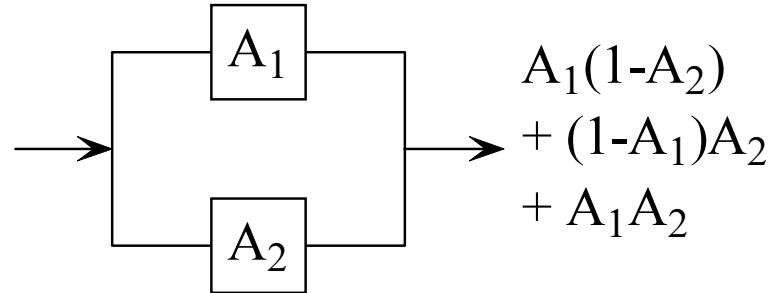
where MTBF = the mean time between failures and MTTR = the mean time to repair.

The availability of a system depends upon the availability of its individual components and the system organisation. For example the system may be organised to make use of redundant components in the case component failure or it may still be able to function but with reduced capabilities.

Availability of serial connections



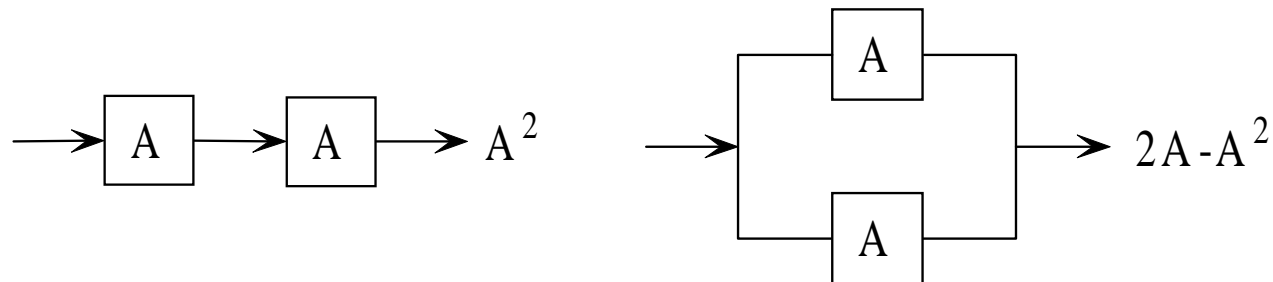
Availability of parallel connections



$A_1(1 - A_2)$ = the probability that link 1 is up times the probability that link 2 is down

$(1 - A_1)A_2$ = the probability that link 1 is down times the probability that link 2 is up

A_1A_2 = the probability that link 1 is up times the probability that link 2 is up



Examples

Two sites are connected via a link between two remote bridges. If the availability of each bridge is 0.97 then the availability of the link between the sites is $0.97 \times 0.97 = 0.9409$. This assumes the availability of the communications link between the bridges is 1. If it is 0.999 the overall availability is $0.97 \times 0.97 \times 0.999 = 0.93996$.

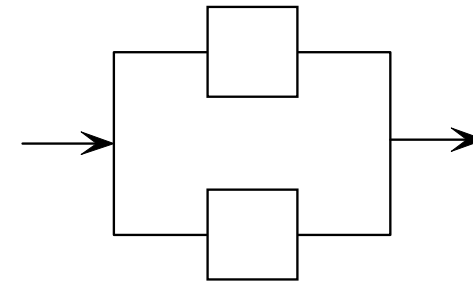
If a terminal is connected to a host via two links in such a way that should one link fail the other is automatically used for back up and the availability of each link is 0.99 the probability that a link is unavailable is $1 - 0.99 = 0.01$. So the probability that both are unavailable is $0.01 \times 0.01 = 0.0001$. Therefore the availability of the combined link is $1 - 0.0001 = 0.9999$.

Availability and Load

Complex configurations complicate the analysis of system availability as does taking into account the load on the system.

Example

Consider a dual link system as shown on the left in which nonpeak periods account for 30% of requests. During nonpeak periods either link can handle the traffic load. During peak periods both links are required to handle the full load, but a single link can only handle 75% of the peak. If the availability of either link is 0.92, on average what percentage of requests for service can be handled by the system.



The functional availability, $A_f = (\text{capacity of one link}) \times (\text{probability one link is up})$
 $+ (\text{capacity of both links}) \times (\text{probability both links are up})$

The probability that both links are up is A^2 , where A is the availability of either link. The probability that exactly one link is up is $A(1 - A) + (1 - A)A = 2A - 2A^2$. So in this case the probability that both links are up is $0.92^2 = 0.8464$ and the probability that exactly one link is up is $2 \times 0.92 - 2 \times 0.92^2 = 0.1472$.

Since one link is sufficient for nonpeak loads

$$A_f(\text{nonpeak}) = (1.0) \times (0.1472) + (1.0) \times (0.8464) = 0.9936$$

and, for peak periods,

$$A_f(\text{peak}) = (0.75) \times (0.1472) + (1.0) \times (0.8464) = 0.9567$$

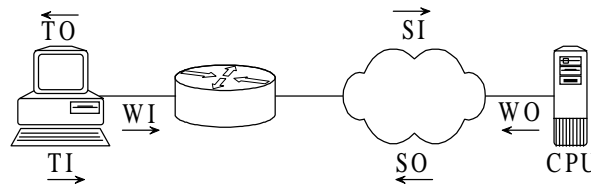
The overall functional availability is

$$\begin{aligned} A_f &= 0.7 \times A_f(\text{peak}) + 0.3 \times A_f(\text{nonpeak}) \\ &= 0.7 \times 0.9567 + 0.3 \times 0.9936 \\ &= 0.9678 \end{aligned}$$

So on average 97% of requests can be handled by the system.

Response Time (Latency)

The overall response time in a network is of little value. The overall response time is made up of several elements.



$$RT = TI + WI + SI + CPU + WO + SO + TO$$

RT = response time

CPU = CPU processor delay

TI = inbound terminal delay

WO = outbound queuing time

WI = inbound queuing time

SO = outbound service time

SI = inbound service time

TO = outbound terminal delay

Accuracy

Due to the built-in error-correction mechanisms in protocols in the data-link and transport layers, accuracy is not usually a concern. It is, however, useful to monitor the rate of errors. Error rates may indicate a source of noise or interference that should be corrected.

Efficiency-oriented Performance Indicators

Throughput

Throughput metrics are measured in units of inverse time. For example:

- transactions completed per minute

- gigabytes of data written to tape per hour

- memory accesses per second

- megabits of data transmitted per second.

“Bandwidth” is often erroneously used to describe the theoretical maximum throughput of a data link or other device. Capacity is a better description of the theoretical maximum throughput.

Efficiency is defined as the ratio of usable throughput to the theoretical maximum throughput. For computer networks, where packets may be lost or damaged, the term goodput (Ugh!!!) is sometimes used for the arrival rate of undamaged packets.

Utilization

The fraction of time that a system component, such as CPU, disk, or data link, is active is its utilisation. It follows from this definition that utilisation values range between 0 and 1. The maximum throughput of a system is reached when the busiest component reaches a utilisation of 1. Response time increases rapidly as utilisation approaches 100%, so that many systems are designed to keep utilisation below some threshold such as 70% or 80%.